PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-251197
(43)Date of publication of application : 06.09.2002

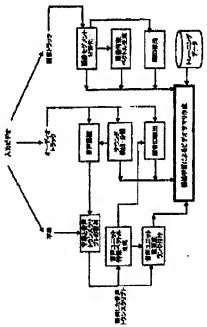
(51)Int.Cl. G10L 15/00
G10L 15/06
G10L 15/10
G10L 15/16
G10L 17/00
H04N 5/91

(21)Application number : 2001-376561 (71)Applicant : NEC CORP
(22)Date of filing : 11.12.2001 (72)Inventor : IKOU KYOU
LIU XIN

(30)Priority
Priority number : 2000 254534 Priority date : 12.12.2000 Priority country : US
2001 011215 25.10.2001 US

(54) AUDIOVISUAL SUMMARY CREATING METHOD

(57)Abstract:
PROBLEM TO BE SOLVED: To create an audio-centric, an image-centric, and an integrated audiovisual summaries of high quality by seamlessly integrating an image, audio, and text features extracted from input video.
SOLUTION: Integrated summarization is employed when strict synchronization of audio content and an image content is not required. A video programming which requires synchronization of the audio content and the image content is summarized by using an audio-centric approach or an image-centric approach. Both a machine learning-based approach and an alternative, heuristics-based approach are usable. Various probabilistic methods such as a naive Bayes method, a decision tree method, a neural network method, and a maximum entropy method are employed with the machine learning-based learning approach. To create an integrated audiovisual summary by using the alternative, heuristics-based approach, a maximum-bipartite-matching approach is employed.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2002－251197
（P2002－251197Ａ）
(43)公開日　平成14年９月６日(2002.9.6)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード（参考）
Ｇ１０Ｌ　15/00		Ｇ１０Ｌ　3/00	５５１Ｇ	５Ｃ０５３
15/06			５２１Ｆ	５Ｄ０１５
15/10			５３１Ｎ	
15/16			５３９	
17/00			５４５Ａ	

審査請求　有　　請求項の数106　ＯＬ　（全 19 頁）　最終頁に続く

(21)出願番号　特願2001－376561(P2001－376561)

(22)出願日　平成13年12月11日(2001.12.11)

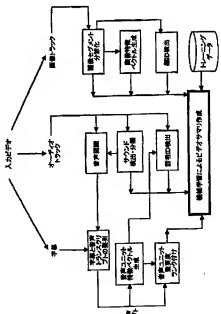
(31)優先権主張番号　６０／２５４５３４
(32)優先日　平成12年12月12日(2000.12.12)
(33)優先権主張国　米国（ＵＳ）
(31)優先権主張番号　１０／０１１２１５
(32)優先日　平成13年10月25日(2001.10.25)
(33)優先権主張国　米国（ＵＳ）

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(72)発明者　イコウ　キョウ
アメリカ合衆国，ニュージャージー
08540　プリンストン，４　インディペンデンス　ウエイ，エヌ・イー・シー・ユー・エス・エー　インク内
(74)代理人　100097157
弁理士　桂木　雄二

最終頁に続く

(54)【発明の名称】　オーディオビジュアルサマリ作成方法

(57)【要約】
【課題】　入力ビデオから抽出された画像、オーディオ、およびテキスト特徴をシームレスに統合することにより、オーディオ中心型、画像中心型、およびオーディオビジュアル統合型の高品質のサマリを作成する。
【解決手段】　オーディオと画像の内容の厳密な同期が要求されないときには、統合型サマリ作成が用いられる。オーディオ内容と画像内容の同期を要求するビデオ番組の場合、オーディオ中心型または画像中心型のいずれかの方法を用いてサマリが作成される。機械学習による方法と、代替法である発見的方法が使用可能である。ナイーブベイズ法、決定木法、ニューラルネットワーク法、および最大エントロピー法のようなさまざまな確率論的方法が、機械学習による方法で使用可能である。代替法である発見的方法を用いてオーディオビジュアル統合型サマリを作成するには、最大２部マッチング法が用いられる。



【特許請求の範囲】
【請求項１】　オーディオトラックおよび画像トラックを有するビデオ番組のオーディオ中心型オーディオビジュアルサマリを作成する方法において、
前記オーディオビジュアルサマリの時間長Ｌｓｕｍ　を選択するステップと、
前記オーディオトラックおよび画像トラックを検査するステップと、
前記オーディオビジュアルサマリの所望される内容に関連する１個以上の所定のオーディオ、画像、およびテキスト特性に基づいて、前記オーディオトラックから１個以上のオーディオセグメントを識別し、当該識別が、前記ビデオ番組内のオーディオセグメントのそれぞれについて、与えられたオーディオセグメントが前記オーディオビジュアルサマリに含められるのに適している確率を与える、前もって生成された経験に基づく学習データに依拠する機械学習法に従って実行される識別ステップと、
前記オーディオセグメントを前記オーディオビジュアルサマリに追加するステップと、
時間長Ｌｓｕｍ　に達するまで、前記確率の降順に前記識別および追加を実行するステップと、
１個以上の識別されたオーディオセグメントに対応する１個以上の画像セグメントのみを、前記１個以上のオーディオセグメントと前記１個以上の画像セグメントの間の同期の程度が高くなるように、選択するステップとを有することを特徴とするオーディオビジュアルサマリ作成方法。
【請求項２】　前記識別するステップは、
非音声サウンドを含むオーディオセグメントを検出するステップと、
内容に従って前記非音声サウンドを分類するステップと、
前記非音声サウンドのそれぞれについて、開始時刻コード、長さ、およびカテゴリを出力するステップとを有することを特徴とする請求項１記載の方法。
【請求項３】　前記オーディオセグメントが音声を含むとき、前記識別するステップは、
前記オーディオセグメントに対する音声認識を実行して音声トランスクリプトを生成するステップと、
前記音声トランスクリプトのそれぞれについて、開始時刻コードおよび長さを出力するステップとを有することを特徴とする請求項２記載の方法。
【請求項４】　字幕が存在するとき、前記方法は、字幕と音声トランスクリプトを整列させるステップをさらに有することを特徴とする請求項３記載の方法。
【請求項５】　前記識別するステップは、
前記字幕が存在する場合には前記整列に基づいて、また、前記字幕が存在しない場合には前記音声トランスクリプトに基づいて、音声ユニットを生成するステップと、
前記音声ユニットのそれぞれについて、特徴ベクトルを生成するステップとを有することを特徴とする請求項４記載の方法。
【請求項６】　前記音声ユニットのそれぞれについて、重要度ランクを計算するステップをさらに有することを特徴とする請求項５記載の方法。
【請求項７】　前記音声ユニットを受け取るステップと、
１以上の話者の識別を決定するステップとをさらに有することを特徴とする請求項６記載の方法。
【請求項８】　前記識別するステップは、前記画像トラックを個々の画像セグメントに分節化するステップを有することを特徴とする請求項１記載の方法。
【請求項９】　画像特徴を抽出するステップと、
前記画像セグメントのそれぞれについて、画像特徴ベクトルを形成するステップとをさらに有することを特徴とする請求項８記載の方法。
【請求項１０】　前記画像セグメントのそれぞれについて、１個以上の顔の識別を決定するステップをさらに有することを特徴とする請求項９記載の方法。
【請求項１１】　前記確率は、ナイーブベイズ法、決定木法、ニューラルネットワーク法、および最大エントロピー法からなる群から選択される方法に従って計算されることを特徴とする請求項１記載の方法。
【請求項１２】　オーディオトラックおよび画像トラックを有するビデオ番組の画像中心型オーディオビジュアルサマリを作成する方法において、
前記オーディオビジュアルサマリの時間長Ｌｓｕｍ　を選択するステップと、
前記ビデオ番組の前記画像トラックおよびオーディオトラックを検査するステップと、
前記オーディオビジュアルサマリの所望される内容に関連する１個以上の所定の画像、オーディオ、およびテキスト特性に基づいて、前記画像トラックから１個以上の画像セグメントを識別し、当該識別が、前記ビデオ番組内の前記画像セグメントのそれぞれについて、与えられた画像セグメントが前記オーディオビジュアルサマリに含められるのに適している確率を与える、前もって生成された経験に基づく学習データに依拠する機械学習法に従って実行される識別ステップと、
前記１個以上の画像セグメントを前記オーディオビジュアルサマリに追加するステップと、
時間長Ｌｓｕｍ　に達するまで、前記確率の降順に前記識別および追加を実行するステップと、
１個以上の識別された画像セグメントに対応する１個以上のオーディオセグメントのみを、前記１個以上の画像セグメントと前記１個以上のオーディオセグメントの間の同期の程度が高くなるように、選択するステップとを有することを特徴とするオーディオビジュアルサマリ作

成方法。

【請求項１３】　前記識別するステップは、前記画像トラックを個々の画像セグメントに分節化するステップを有することを特徴とする請求項１２記載の方法。

【請求項１４】　画像特徴を抽出するステップと、前記画像セグメントのそれぞれについて、画像特徴ベクトルを形成するステップとをさらに有することを特徴とする請求項１３記載の方法。

【請求項１５】　前記画像セグメントのそれぞれについて、１個以上の顔の識別を決定するステップをさらに有することを特徴とする請求項１０記載の方法。

【請求項１６】　前記オーディオビジュアルサマリ内の前記画像セグメントのそれぞれについて、最小再生時間L_{min}を選択するステップをさらに有することを特徴とする請求項１２記載の方法。

【請求項１７】　比較的多数のオーディオセグメントおよび画像セグメントが前記オーディオビジュアルサマリに提供されて、幅指向のオーディオビジュアルサマリを提供するように、L_{min}はL_{sum}に比べて十分に小さいことを特徴とする請求項１６記載の方法。

【請求項１８】　比較的少数のオーディオセグメントおよび画像セグメントが前記オーディオビジュアルサマリに提供されて、深さ指向のオーディオビジュアルサマリを提供するように、L_{min}はL_{sum}に比べて十分に大きいことを特徴とする請求項１６記載の方法。

【請求項１９】　前記識別するステップは、
非音声サウンドを含むオーディオセグメントを検出するステップと、
内容に従って前記非音声サウンドを分類するステップと、
前記非音声サウンドのそれぞれについて、開始時刻コード、長さ、およびカテゴリを出力するステップとを有することを特徴とする請求項１２記載の方法。

【請求項２０】　前記オーディオセグメントが音声を含むとき、前記識別するステップは、
前記オーディオセグメントに対する音声認識を実行して音声トランスクリプトを生成するステップと、
前記音声トランスクリプトのそれぞれについて、開始時刻コードおよび長さを出力するステップとを有することを特徴とする請求項１９記載の方法。

【請求項２１】　字幕が存在するとき、前記方法は、字幕と音声トランスクリプトを整列させるステップをさらに有することを特徴とする請求項２０記載の方法。

【請求項２２】　前記識別するステップは、
前記字幕が存在する場合には前記整列に基づいて、また、前記字幕が存在しない場合には前記音声トランスクリプトに基づいて、音声ユニットを生成するステップと、
前記音声ユニットのそれぞれについて、特徴ベクトルを生成するステップとを有することを特徴とする請求項１記載の方法。

【請求項２３】　前記音声ユニットのそれぞれについて、重要度ランクを計算するステップをさらに有することを特徴とする請求項２２記載の方法。

【請求項２４】　前記音声ユニットを受け取るステップと、
１以上の話者の識別を決定するステップとをさらに有することを特徴とする請求項２３記載の方法。

【請求項２５】　前記確率は、ナイーブベイズ法、決定木法、ニューラルネットワーク法、および最大エントロピー法からなる群から選択される方法に従って計算されることを特徴とする請求項１２記載の方法。

【請求項２６】　オーディオトラックおよびビデオトラックを有するビデオ番組の統合オーディオビジュアルサマリを作成する方法において、
前記オーディオビジュアルサマリの時間長L_{sum}を選択するステップと、
オーディオビジュアルサマリに含まれるべき前記画像セグメントのそれぞれについて、最小再生時間L_{min}を選択するステップと、
前記オーディオビジュアルサマリの長さL_{sum}に達するまで１個以上の所望されるオーディオセグメントを選択し、当該選択が、前記ビデオ番組内の前記オーディオセグメントのそれぞれについて、与えられたオーディオセグメントが前記オーディオビジュアルサマリに含められるのに適している確率を与える、前もって生成された経験に基づく学習データに依拠する機械学習法に従って実行されることによりオーディオサマリを作成するステップと、
前記画像セグメントのそれぞれについて、前記機械学習法に従って、与えられた画像セグメントが前記オーディオビジュアルサマリに含められるのに適している確率を計算するステップと、
選択された前記オーディオセグメントのそれぞれについて、対応する画像セグメントに関して、得られるオーディオセグメントと画像セグメントのペアが、あらかじめ規定された整列要件を満たすかどうかを検査するステップと、
得られるオーディオセグメントと画像セグメントのペアが、あらかじめ規定された整列要件を満たす場合、ペアのオーディオセグメントと画像セグメントをそれぞれの最初から前記最小再生時間L_{min}だけ整列させて第１の整列点を規定するステップと、
前記検査および識別を繰り返して前記整列点をすべて識別するステップと、
前記オーディオビジュアルサマリの全長を、
前記オーディオビジュアルサマリの最初から開始し第１の整列点で終了するか、
ある整列点における画像セグメントの最後から開始し次の整列点で終了するか、

最後の整列点における画像セグメントの最後から開始し前記オーディオビジュアルサマリの最後で終了するかのいずれかの期間をそれぞれ有する複数のパーティションに分割するステップと、
前記パーティションのそれぞれについて、以下のステップ、すなわち、
該パーティションの期間に入る画像セグメントの集合を識別するステップと、
前記パーティションに挿入されることが可能な画像セグメントの個数を決定するステップと、
挿入されるべきと識別された画像セグメントの長さを決定するステップと、
与えられた画像セグメントが前記オーディオビジュアルサマリに含められるのに適している前記確率の降順に、識別された画像セグメントを前記個数だけ選択するステップと、
選択された画像セグメントのそれぞれについて、それぞれの最初から前記時間長だけのセクションを収集し、すべての収集されたセクションを時間の降順に前記パーティションに追加するステップとに従って、さらに画像セグメントを追加するステップとを有することを特徴とするオーディオビジュアルサマリ作成方法。

【請求項２７】　前記識別するステップは、
非音声サウンドを含むオーディオセグメントを検出するステップと、
内容に従って前記非音声サウンドを分類するステップと、
前記非音声サウンドのそれぞれについて、開始時刻コード、長さ、およびカテゴリを出力するステップとを有することを特徴とする請求項２６記載の方法。

【請求項２８】　前記オーディオセグメントが音声を含むとき、前記識別するステップは、
前記オーディオセグメントに対する音声認識を実行して音声トランスクリプトを生成するステップと、
前記音声トランスクリプトのそれぞれについて、開始時刻コードおよび長さを出力するステップとを有することを特徴とする請求項２７記載の方法。

【請求項２９】　字幕が存在するとき、前記方法は、字幕と音声トランスクリプトを整列させるステップをさらに有することを特徴とする請求項２８記載の方法。

【請求項３０】　前記字幕が存在する場合には前記整列に基づいて、また、前記字幕が存在しない場合には前記音声トランスクリプトに基づいて、音声ユニットを生成するステップと、
前記音声ユニットのそれぞれについて、特徴ベクトルを生成するステップとをさらに有することを特徴とする請求項２９記載の方法。

【請求項３１】　前記音声ユニットのそれぞれについて、重要度ランクを計算するステップをさらに有することを特徴とする請求項３０記載の方法。

【請求項３２】　前記音声ユニットを受け取るステップと、
１以上の話者の識別を決定するステップとをさらに有することを特徴とする請求項３１記載の方法。

【請求項３３】　比較的多数の画像セグメントが前記オーディオビジュアルサマリに提供されて、幅指向のオーディオビジュアルサマリを提供するように、L_{wv} は L_{sm} に比べて十分に小さいことを特徴とする請求項２６記載の方法。

【請求項３４】　比較的少数の画像セグメントが前記オーディオビジュアルサマリに提供されて、深さ指向のオーディオビジュアルサマリを提供するように、L_{wv} は L_{sm} に比べて十分に大きいことを特徴とする請求項２６記載の方法。

【請求項３５】　前記与えられたオーディオセグメントが前記オーディオビジュアルサマリに含められるのに適している確率は、ナイーブベイズ法、決定木法、ニューラルネットワーク法、および最大エントロピー法からなる群から選択される方法に従って計算されることを特徴とする請求項２６記載の方法。

【請求項３６】　前記与えられた画像セグメントが前記オーディオビジュアルサマリに含められるのに適している確率は、ナイーブベイズ法、決定木法、ニューラルネットワーク法、および最大エントロピー法からなる群から選択される方法に従って計算されることを特徴とする請求項２６記載の方法。

【請求項３７】　前記識別するステップは、前記画像トラックを個々の画像セグメントに分節化するステップを有することを特徴とする請求項２６記載の方法。

【請求項３８】　画像特徴を抽出するステップと、
前記画像セグメントのそれぞれについて、画像特徴ベクトルを形成するステップとをさらに有することを特徴とする請求項３７記載の方法。

【請求項３９】　前記画像セグメントのそれぞれについて、１個以上の顔の識別を決定するステップをさらに有することを特徴とする請求項３８記載の方法。

【請求項４０】　オーディオトラックおよび画像トラックを有するビデオ番組のオーディオ中心型オーディオビジュアルサマリを作成する方法において、
前記オーディオビジュアルサマリの時間長 L_{sm} を選択するステップと、
前記オーディオトラックおよび画像トラックを検査するステップと、
前記オーディオビジュアルサマリの所望される内容に関連する１個以上の所定のオーディオ、画像、およびテキスト特性に基づいて、前記オーディオトラックから１個以上のオーディオセグメントを識別し、当該識別が、前記ビデオ番組内の前記オーディオセグメントのそれぞれについて、与えられたオーディオセグメントが前記オーディオビジュアルサマリに含められるのに適しているか

どうかを決定するランク付けを与える、所定の発見的ルールの集合に従って実行される識別ステップと、
前記オーディオセグメントを前記オーディオビジュアルサマリに追加するステップと、
時間長L_{sum}に達するまで、前記オーディオセグメントのランク付けの降順に前記識別および追加を実行するステップと、
１個以上の識別されたオーディオセグメントに対応する１個以上の画像セグメントのみを、前記１個以上のオーディオセグメントと前記１個以上の画像セグメントの間の同期の程度が高くなるように、選択するステップとを有することを特徴とするオーディオビジュアルサマリ作成方法。

【請求項４１】　前記識別するステップは、
非音声サウンドを含むオーディオセグメントを検出するステップと、
内容に従って前記非音声サウンドを分類するステップと、
前記非音声サウンドのそれぞれについて、開始時刻コード、長さ、およびカテゴリを出力するステップとを有することを特徴とする請求項４０記載の方法。

【請求項４２】　前記オーディオセグメントが音声を含むとき、前記識別するステップは、
前記オーディオセグメントに対する音声認識を実行して音声トランスクリプトを生成するステップと、
前記音声トランスクリプトのそれぞれについて、開始時刻コードおよび長さを出力するステップとを有することを特徴とする請求項４１記載の方法。

【請求項４３】　字幕が存在するとき、前記方法は、字幕と音声トランスクリプトを整列させるステップをさらに有することを特徴とする請求項４２記載の方法。

【請求項４４】　前記字幕が存在する場合には前記整列に基づいて、また、前記字幕が存在しない場合には前記音声トランスクリプトに基づいて、音声ユニットを生成するステップと、
前記音声ユニットのそれぞれについて、特徴ベクトルを生成するステップとをさらに有することを特徴とする請求項４３記載の方法。

【請求項４５】　前記音声ユニットを受け取るステップと、
１以上の話者の識別を決定するステップとをさらに有することを特徴とする請求項４４記載の方法。

【請求項４６】　前記識別するステップは、前記画像トラックを個々の画像セグメントに分節化するステップを有することを特徴とする請求項４０記載の方法。

【請求項４７】　画像特徴を抽出するステップと、
前記画像セグメントのそれぞれについて、画像特徴ベクトルを形成するステップとをさらに有することを特徴とする請求項４６記載の方法。

【請求項４８】　前記画像セグメントのそれぞれについて、１個以上の顔の識別を決定するステップをさらに有することを特徴とする請求項４７記載の方法。

【請求項４９】　前記音声ユニットのそれぞれについて前記ランク付けを計算するステップをさらに有することを特徴とする請求項４０記載の方法。

【請求項５０】　オーディオトラックおよび画像トラックを有するビデオ番組の画像中心型オーディオビジュアルサマリを作成する方法において、
前記サマリの時間長L_{sum}を選択するステップと、
前記画像トラックおよびオーディオトラックを検査するステップと、
前記オーディオビジュアルサマリの所望される内容に関連する１個以上の所定の画像、オーディオ、およびテキスト特性に基づいて、前記画像トラックから１個以上の画像セグメントを識別し、当該識別が、前記ビデオ番組内の前記画像セグメントのそれぞれについて、与えられた画像セグメントが前記オーディオビジュアルサマリに含められるのに適しているかどうかを決定するランク付けを与える、所定の発見的ルールの集合に従って実行される識別ステップと、
前記１個以上の画像セグメントを前記オーディオビジュアルサマリに追加するステップと、
時間長L_{sum}に達するまで、前記ランク付けの降順に前記識別および追加を実行するステップと、
１個以上の識別された画像セグメントに対応する１個以上のオーディオセグメントのみを、前記１個以上の画像セグメントと前記１個以上のオーディオセグメントの間の同期の程度が高くなるように、選択するステップとを有することを特徴とするオーディオビジュアルサマリ作成方法。

【請求項５１】　前記識別するステップは、所定の視覚的類似性および動的特性に基づいて、前記ビデオ番組の画像セグメントをクラスタ化するステップを有することを特徴とする請求項５０記載の方法。

【請求項５２】　前記識別するステップは、前記画像トラックを個々の画像セグメントに分節化するステップを有することを特徴とする請求項５１記載の方法。

【請求項５３】　画像特徴を抽出するステップと、
前記フレームクラスタのそれぞれについて、画像特徴ベクトルを形成するステップとをさらに有することを特徴とする請求項５２記載の方法。

【請求項５４】　前記フレームクラスタのそれぞれについて、１個以上の顔の識別を決定するステップをさらに有することを特徴とする請求項５３記載の方法。

【請求項５５】　前記識別するステップは、
非音声サウンドを含むオーディオセグメントを検出するステップと、
内容に従って前記非音声サウンドを分類するステップと、
前記非音声サウンドのそれぞれについて、開始時刻コー

ド、長さ、およびカテゴリを出力するステップとを有することを特徴とする請求項５０記載の方法。

【請求項５６】　前記オーディオセグメントが音声を含むとき、前記識別するステップは、

前記オーディオセグメントに対する音声認識を実行して音声トランスクリプトを生成するステップと、

前記音声トランスクリプトのそれぞれについて、開始時刻コードおよび長さを出力するステップとを有することを特徴とする請求項５５記載の方法。

【請求項５７】　字幕が存在するとき、前記方法は、字幕と音声トランスクリプトを整列させるステップをさらに有することを特徴とする請求項５６記載の方法。

【請求項５８】　前記字幕が存在する場合には前記整列に基づいて、また、前記字幕が存在しない場合には前記音声トランスクリプトに基づいて、音声ユニットを生成するステップと、

前記音声ユニットのそれぞれについて、特徴ベクトルを生成するステップとをさらに有することを特徴とする請求項５７記載の方法。

【請求項５９】　前記音声ユニットのそれぞれについて、重要度ランクを計算するステップをさらに有することを特徴とする請求項５８記載の方法。

【請求項６０】　前記音声ユニットを受け取るステップと、

１以上の話者の識別を決定するステップとをさらに有することを特徴とする請求項５９記載の方法。

【請求項６１】　前記オーディオビジュアルサマリ内の前記画像セグメントのそれぞれについて、最小再生時間L_{min}を選択するステップをさらに有することを特徴とする請求項５０記載の方法。

【請求項６２】　比較的多数のオーディオセグメントおよび画像セグメントが前記オーディオビジュアルサマリに提供されて、幅指向のオーディオビジュアルサマリを提供するように、L_{min}はL_{sum}に比べて十分に小さいことを特徴とする請求項６１記載の方法。

【請求項６３】　比較的少数のオーディオセグメントおよび画像セグメントが前記オーディオビジュアルサマリに提供されて、深さ指向のオーディオビジュアルサマリを提供するように、L_{min}はL_{sum}に比べて十分に大きいことを特徴とする請求項６１記載の方法。

【請求項６４】　オーディオトラックおよびビデオトラックを有するビデオ番組の統合オーディオビジュアルサマリを作成する方法において、

前記オーディオビジュアルサマリの長さL_{sum}を選択するステップと、

オーディオビジュアルサマリに含まれるべき複数の画像セグメントのそれぞれについて、最小再生時間L_{min}を選択するステップと、

前記ビデオ番組内の前記オーディオセグメントのそれぞれについて、与えられたオーディオセグメントが前記オーディオビジュアルサマリに含められるのに適しているかどうかを決定するランク付けを与える、所定の発見的ルールの集合に従って、１個以上の所望されるオーディオセグメントを選択することによって、オーディオサマリを作成するステップと、

前記オーディオビジュアルサマリの長さに達するまで、前記オーディオセグメントのランク付けの降順に、前記選択を実行するステップと、

各フレームクラスタが少なくとも１つの前記画像セグメントを含み、与えられたフレームクラスタ内のすべての画像セグメントが互いに視覚的に類似しているように、前記画像セグメントの視覚的類似性および動的特性に基づいて、前記ビデオ番組の前記画像セグメントを複数のフレームクラスタへとグループ分けするステップと、

選択された前記オーディオセグメントのそれぞれについて、対応する画像セグメントに関して、得られるオーディオセグメントと画像セグメントのペアが、あらかじめ規定された整列要件を満たすかどうかを検査するステップと、

得られるオーディオセグメントと画像セグメントのペアが、あらかじめ規定された整列要件を満たす場合、ペアのオーディオセグメントと画像セグメントをそれぞれの最初から前記最小再生時間L_{min}だけ整列させて第１の整列点を規定するステップと、

前記検査および識別を繰り返して前記整列点をすべて識別するステップと、

前記オーディオビジュアルサマリの全長を、

前記オーディオビジュアルサマリの最初から開始し第１の整列点で終了するか、

ある整列点における画像セグメントの最後から開始し次の整列点で終了するか、

最後の整列点における画像セグメントの最後から開始し前記オーディオビジュアルサマリの最後で終了するかのいずれかの期間をそれぞれ有する複数のパーティションに分割するステップと、

各時間スロットが前記最小再生時間L_{min}に等しい長さを有するように、前記パーティションのそれぞれを複数の時間スロットに分割するステップと、

前記フレームクラスタと前記時間スロットの間の最適マッチングに従って、以下のこと、すなわち、

各フレームクラスタをただ１つの時間スロットに割り当てること、および、オーディオビジュアルサマリ内のすべての画像セグメントの時間順序を維持することに基づいて、前記パーティションのそれぞれの前記時間スロットを満たすように前記フレームクラスタを割り当てるステップとを有することを特徴とするオーディオビジュアルサマリ作成方法。

【請求項６５】　前記最適マッチングは、最大２部マッチング法によって計算されることを特徴とする請求項６４記載の方法。

【請求項６６】　フレームクラスタより多くの時間スロットがある場合、複数の画像セグメントを含むフレームクラスタを識別し、前記オーディオビジュアルサマリ内の前記画像セグメントの時間順序を維持しながら、すべての前記時間スロットが満たされるまで、前記識別されたフレームクラスタからの画像セグメントを時間スロットに割り当てることを特徴とする請求項６５記載の方法。

【請求項６７】　前記時間順序が維持されていることを確認するために前記オーディオビジュアルサマリを検査するステップと、

前記時間順序が維持されていない場合、前記時間順序が維持されるように、各パーティションに追加された前記画像セグメントを並べ替えるステップとをさらに有することを特徴とする請求項６６記載の方法。

【請求項６８】　前記識別するステップは、

非音声サウンドを含むオーディオセグメントを検出するステップと、

内容に従って前記非音声サウンドを分類するステップと、

前記非音声サウンドのそれぞれについて、開始時刻コード、長さ、およびカテゴリを出力するステップとを有することを特徴とする請求項６４記載の方法。

【請求項６９】　前記オーディオセグメントが音声を含むとき、前記識別するステップは、

前記オーディオセグメントに対する音声認識を実行して音声トランスクリプトを生成するステップと、

前記音声トランスクリプトのそれぞれについて、開始時刻コードおよび長さを出力するステップとを有することを特徴とする請求項６８記載の方法。

【請求項７０】　字幕が存在するとき、前記方法は、字幕と音声トランスクリプトを整列させるステップをさらに有することを特徴とする請求項６９記載の方法。

【請求項７１】　前記字幕が存在する場合には前記整列に基づいて、また、前記字幕が存在しない場合には前記音声トランスクリプトに基づいて、音声ユニットを生成するステップと、

前記音声ユニットのそれぞれについて、特徴ベクトルを生成するステップとをさらに有することを特徴とする請求項７０記載の方法。

【請求項７２】　前記音声ユニットのそれぞれについて、重要度ランクを計算するステップをさらに有することを特徴とする請求項７１記載の方法。

【請求項７３】　前記音声ユニットを受け取るステップと、

１以上の話者の識別を決定するステップとをさらに有することを特徴とする請求項７２記載の方法。

【請求項７４】　比較的多数の画像セグメントが前記オーディオビジュアルサマリに提供されて、幅指向のオーディオビジュアルサマリを提供するように、L_{min} は L_{max} に比べて十分に小さいことを特徴とする請求項６４記載の方法。

【請求項７５】　比較的少数の画像セグメントが前記オーディオビジュアルサマリに提供されて、深さ指向のオーディオビジュアルサマリを提供するように、L_{min} は L_{max} に比べて十分に大きいことを特徴とする請求項６４記載の方法。

【請求項７６】　前記識別するステップは、前記画像トラックを個々の画像セグメントに分節化するステップを有することを特徴とする請求項６４記載の方法。

【請求項７７】　画像特徴を抽出するステップと、

前記フレームクラスタのそれぞれについて、画像特徴ベクトルを形成するステップとをさらに有することを特徴とする請求項７６記載の方法。

【請求項７８】　前記画像セグメントのそれぞれについて、１個以上の顔の識別を決定するステップをさらに有することを特徴とする請求項７７記載の方法。

【請求項７９】　オーディオビジュアルコンテンツからなるビデオ番組のビデオサマリを作成する装置において、

前記オーディオビジュアルコンテンツのオーディオトラックおよび画像トラックを検査する検査手段と、

前記ビデオサマリの所望されるコンテンツに関連する所定のオーディオ、画像、およびテキスト特性のうちの少なくとも１つに基づき、前記オーディオトラックから１個以上のオーディオセグメントを、前記画像トラックから１個以上の画像セグメントを、前記ビデオサマリに含められるのに適しているかどうかを決定する順位を与える所定の基準に従って、それぞれ識別する手段と、

前記順位に従って、前記１個以上のオーディオセグメントおよび前記１個以上の画像セグメントをそれぞれ時間軸上に配置して前記ビデオサマリを生成する手段と、

を有することを特徴とするビデオサマリ作成装置。

【請求項８０】　前記識別する手段は、

前記ビデオサマリの所望される内容に関連する１個以上の所定のオーディオ、画像、およびテキスト特性に基づいて、前記オーディオトラックから１個以上のオーディオセグメントを識別する際に、当該識別を、前記ビデオ番組内のオーディオセグメントのそれぞれについて、与えられたオーディオセグメントが前記オーディオビジュアルサマリに含められるのに適している確率を与える、前もって生成された経験に基づく学習データに依拠する機械学習法に従って実行する、ことを特徴とする請求項７９記載のビデオサマリ作成装置。

【請求項８１】　前記識別する手段は、前記オーディオセグメントをカテゴリ化することを特徴とする請求項８０記載のビデオサマリ作成装置。

【請求項８２】　前記オーディオセグメントは、音声と非音声とにカテゴリ化されることを特徴とする請求項８１記載のビデオサマリ作成装置。

13

【請求項８３】　前記識別する手段は、
非音声サウンドを含むオーディオセグメントを検出し、内容に従って前記非音声サウンドを分類し、前記非音声サウンドのそれぞれについて、オーディオ情報を出力することを特徴とする請求項８２記載のビデオサマリ作成装置。
【請求項８４】　前記オーディオ情報は、開始時刻コード、長さ、およびカテゴリであることを特徴とする請求項８３記載のビデオサマリ作成装置。
【請求項８５】　前記識別する手段は、
前記オーディオビジュアルサマリの所望される内容に関連する１個以上の所定の画像、オーディオ、およびテキスト特性に基づいて、前記画像トラックから１個以上の画像セグメントを識別し、当該識別が、前記ビデオ番組内の前記画像セグメントのそれぞれについて、与えられた画像セグメントが前記オーディオビジュアルサマリに含められるのに適している確率を与える、前もって生成された経験に基づく学習データに依拠する機械学習法に従って実行する、ことを特徴とする請求項７９記載のビデオサマリ作成装置。
【請求項８６】　前記識別する手段は、
１個以上の所望されるオーディオセグメントを選択し、当該選択が、前記ビデオ番組内の前記オーディオセグメントのそれぞれについて、与えられたオーディオセグメントが前記オーディオビジュアルサマリに含められるのに適している確率を与える、前もって生成された経験に基づく学習データに依拠する機械学習法に従って実行する、ことを特徴とする請求項７９記載のビデオサマリ作成装置。
【請求項８７】　前記識別する手段は、
前記オーディオビジュアルサマリの所望される内容に関連する１個以上の所定のオーディオ、画像、およびテキスト特性に基づいて、前記オーディオトラックから１個以上のオーディオセグメントを識別し、当該識別が、前記ビデオ番組内の前記オーディオセグメントのそれぞれについて、与えられたオーディオセグメントが前記オーディオビジュアルサマリに含められるのに適しているかどうかを決定するランク付けを与える、所定の発見的ルールの集合に従って実行する、ことを特徴とする請求項７９記載のビデオサマリ作成装置。
【請求項８８】　前記識別する手段は、前記オーディオセグメントをカテゴリ化することを特徴とする請求項８７記載のビデオサマリ作成装置。
【請求項８９】　前記オーディオセグメントは、音声と非音声とにカテゴリ化されることを特徴とする請求項８８記載のビデオサマリ作成装置。
【請求項９０】　前記識別する手段は、
非音声サウンドを含むオーディオセグメントを検出し、内容に従って前記非音声サウンドを分類し、前記非音声サウンドのそれぞれについて、オーディオ情報を出力することを特徴とする請求項８９記載のビデオサマリ作成装置。

14

【請求項９１】　前記オーディオ情報は、開始時刻コード、長さ、およびカテゴリであることを特徴とする請求項９０記載のビデオサマリ作成装置。
【請求項９２】　前記識別する手段は、
前記オーディオビジュアルサマリの所望される内容に関連する１個以上の所定の画像、オーディオ、およびテキスト特性に基づいて、前記画像トラックから１個以上の画像セグメントを識別し、当該識別が、前記ビデオ番組内の前記画像セグメントのそれぞれについて、与えられた画像セグメントが前記オーディオビジュアルサマリに含められるのに適しているかどうかを決定するランク付けを与える、所定の発見的ルールの集合に従って実行する、ことを特徴とする請求項７９記載のビデオサマリ作成装置。
【請求項９３】　オーディオビジュアルコンテンツからなるビデオ番組のビデオサマリを作成する方法において、
前記オーディオビジュアルコンテンツのオーディオトラックおよび画像トラックを検査し、
前記ビデオサマリの所望されるコンテンツに関連する所定のオーディオ、画像、およびテキスト特性のうちの少なくとも１つに基づき、前記オーディオトラックから１個以上のオーディオセグメントを、前記画像トラックから１個以上の画像セグメントを、前記ビデオサマリに含められるのに適しているかどうかを決定する順位を与える所定の基準に従って、それぞれ識別し、
前記順位に従って、前記１個以上のオーディオセグメントおよび前記１個以上の画像セグメントをそれぞれ時間軸上に配置して前記ビデオサマリを生成する、
ステップを有することを特徴とするビデオサマリ作成方法。
【請求項９４】　前記識別するステップは、
前記ビデオサマリの所望される内容に関連する１個以上の所定のオーディオ、画像、およびテキスト特性に基づいて、前記オーディオトラックから１個以上のオーディオセグメントを識別する際に、当該識別を、前記ビデオ番組内のオーディオセグメントのそれぞれについて、与えられたオーディオセグメントが前記オーディオビジュアルサマリに含められるのに適している確率を与える、前もって生成された経験に基づく学習データに依拠する機械学習法に従って実行する、ことを特徴とする請求項９３記載のビデオサマリ作成方法。
【請求項９５】　前記識別するステップは、前記オーディオセグメントをカテゴリ化することを特徴とする請求項９４記載のビデオサマリ作成方法。
【請求項９６】　前記オーディオセグメントは、音声と非音声とにカテゴリ化されることを特徴とする請求項９５記載のビデオサマリ作成方法。

【請求項９７】　前記識別するステップは、非音声サウンドを含むオーディオセグメントを検出し、内容に従って前記非音声サウンドを分類し、前記非音声サウンドのそれぞれについて、オーディオ情報を出力することを特徴とする請求項９７記載のビデオサマリ作成方法。

【請求項９８】　前記オーディオ情報は、開始時刻コード、長さ、およびカテゴリであることを特徴とする請求項９７記載のビデオサマリ作成方法。

【請求項９９】　前記識別するステップは、前記オーディオビジュアルサマリの所望される内容に関連する１個以上の所定の画像、オーディオ、およびテキスト特性に基づいて、前記画像トラックから１個以上の画像セグメントを識別し、当該識別が、前記ビデオ番組内の前記画像セグメントのそれぞれについて、与えられた画像セグメントが前記オーディオビジュアルサマリに含められるのに適している確率を与える、前もって生成された経験に基づく学習データに依拠する機械学習法に従って実行する、ことを特徴とする請求項９３記載のビデオサマリ作成方法。

【請求項１００】　前記識別するステップは、１個以上の所望されるオーディオセグメントを選択し、当該選択が、前記ビデオ番組内の前記オーディオセグメントのそれぞれについて、与えられたオーディオセグメントが前記オーディオビジュアルサマリに含められるのに適している確率を与える、前もって生成された経験に基づく学習データに依拠する機械学習法に従って実行する、ことを特徴とする請求項９３記載のビデオサマリ作成方法。

【請求項１０１】　前記識別するステップは、前記オーディオビジュアルサマリの所望される内容に関連する１個以上の所定のオーディオ、画像、およびテキスト特性に基づいて、前記オーディオトラックから１個以上のオーディオセグメントを識別し、当該識別が、前記ビデオ番組内の前記オーディオセグメントのそれぞれについて、与えられたオーディオセグメントが前記オーディオビジュアルサマリに含められるのに適しているかどうかを決定するランク付けを与える、所定の発見的ルールの集合に従って実行する、ことを特徴とする請求項９３記載のビデオサマリ作成方法。

【請求項１０２】　前記識別するステップは、前記オーディオセグメントをカテゴリ化することを特徴とする請求項１０１記載のビデオサマリ作成方法。

【請求項１０３】　前記オーディオセグメントは、音声と非音声とにカテゴリ化されることを特徴とする請求項１０２記載のビデオサマリ作成方法。

【請求項１０４】　前記識別するステップは、非音声サウンドを含むオーディオセグメントを検出し、内容に従って前記非音声サウンドを分類し、前記非音声サウンドのそれぞれについて、オーディオ情報を出力することを特徴とする請求項１０３記載のビデオサマリ作成方法。

【請求項１０５】　前記オーディオ情報は、開始時刻コード、長さ、およびカテゴリであることを特徴とする請求項１０４記載のビデオサマリ作成方法。

【請求項１０６】　前記識別するステップは、前記オーディオビジュアルサマリの所望される内容に関連する１個以上の所定の画像、オーディオ、およびテキスト特性に基づいて、前記画像トラックから１個以上の画像セグメントを識別し、当該識別が、前記ビデオ番組内の前記画像セグメントのそれぞれについて、与えられた画像セグメントが前記オーディオビジュアルサマリに含められるのに適しているかどうかを決定するランク付けを与える、所定の発見的ルールの集合に従って実行する、ことを特徴とする請求項９３記載のビデオサマリ作成方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、一般に、ビデオサマリ作成技術に関し、特に、入力ビデオから抽出した画像、オーディオ、およびテキスト特徴をシームレスに統合することによりビデオサマリを作成する方法およびシステムに関する。

【０００２】

【従来の技術】長い論説や学術論文などのテキスト文書の多くには要約がある。要約の助けにより、読者は、文書全体を詳細に分析せずに、その文書の内容が関心のあるものかどうかをすばやく確かめることができる。テキスト文書の場合もそうであるが、ビデオ番組の内容および性質は一日では捉えられないことが多い。同様にして一般的に全体の内容を示すために、長いビデオ番組の要約すなわちサマリを提供することが一般に望まれる。

【０００３】最近、ワールドワイドウェブ（WWWあるいはウェブ）の爆発的な成長により、オンラインテキストおよびマルチメディアデータコレクションの数が急激に増大している。オンラインマルチメディアコンテンツの増大というこの傾向が続くと、ユーザが大量のデータから最も関連性のある情報をすばやく識別することを支援する自動データサマリ作成技術はますます重要になる。

【０００４】この状況において、ビデオサマリ作成が、困難な課題を提示する。その作業が困難であるのは、ビデオ番組の画像トラックおよびオーディオトラックの両方のサマリ作成をまず必要とするからである。２つのサマリを自然なやりかたで有効に統合することが、もう１つの課題となる。

【０００５】一般に、ほとんどの種類のビデオサマリ作成は、オーディオ中心型サマリ作成（audio-centric summarization）、画像中心型サマリ作成（image-centric summarization）、およびオーディオビジュアル統合型サマ

リ作成(integrated audio-visual summarization)という3つのカテゴリに分類することができる。ビデオ番組のうちには、例えばニュース放送、ドキュメンタリー、ビデオセミナーのように、対応するオーディオトラックと画像トラックの間に強い相関のないタイプのものがある。このようなビデオカテゴリについては、オーディオと画像をゆるく整列（整合）させながら、オーディオと画像の両方の内容のカバレジを最大にするオーディオビジュアル統合型サマリ作成アプローチを使用するのが適当である。他方、映画、ドラマ、トークショーなどのような他のタイプのビデオ番組は、オーディオトラックと画像トラックの間に強い相関を有することがある。この種のビデオ番組については、オーディオ提示とビデオ画像の間の同期が重要である。このような状況では、オーディオ中心型または画像中心型のいずれかのサマリ作成方法を使用するのが適当である。

【0006】

【発明が解決しようとする課題】従来のシステムは、このようなさまざまなタイプのビデオ番組に対する有効で効率的なサマリ作成という課題に対し、包括的な解決法を提供していない。現在使用されている多くのビデオサマリ作成システム・方法は、あるタイプのビデオ内容を発見的に重要であるとみなし、これらのあらかじめ指定した内容を入力ビデオから抽出することによってサマリを作成している。その結果、これらのビデオサマリ作成システム・方法は、非常に領域特異的（領域固有）かつアプリケーション特異的であり、ユーザの個々の需要に基づいてサマリを作成することや、さまざまな種類のビデオ番組を処理することができない。

【0007】

【課題を解決するための手段】本発明は、機械学習フレームワークに基づくビデオサマリ作成のシステムおよび方法を提供することによって、従来のビデオサマリ作成技術の前記およびその他の欠点を克服する。また、本発明はさらに、機械学習フレームワークによって要求されるトレーニングデータを得ることが困難な状況に対処するためのシステムおよび方法も提供する。これらのシステムおよび方法は、入力ビデオから抽出される画像、オーディオ、およびテキスト特徴をシームレスに統合することによって、高品質のオーディオおよび画像のサマリを作成することができる。

【0008】オーディオトラックと画像トラックの間の強い同期に依存しないビデオ番組の具体例として、オーディオセグメントが最近の地震による犠牲者の数に関する情報を提示しているテレビニュース番組を考える。対応する画像セグメントは、現場のレポーターの接写であったり、崩壊した建物の現場で作業する救助隊の接写であったり、地震の震央を示す地域地図の接写であったりする。このような場合、オーディオ内容は、必ずしも、対応する画像内容に言及している必要がないことが多い。前述のように、このようなビデオ番組のその他の例には、ドキュメンタリー、セミナーなどがある。

【0009】本発明の一実施例によれば、厳密な同期が要求されないときには、ビデオ番組のサマリを作成するために、オーディオビジュアル統合型サマリ作成技術が用いられる。このようなビデオ番組のサマリを作成する際には、オーディオおよび画像のサマリを別個に作成することが好ましい。その後、2つのサマリが、ゆるく整列して統合される。このアプローチでは、オーディオ内容と画像内容の両方のカバレジを、サマリにおいて最大化することが可能である。

【0010】逆に、オーディオ内容と画像内容の間の強い同期を要求するビデオ番組は、一般に、特定の瞬間におけるオーディオトラックがその瞬間に提示される画像と直接関係しており、その逆も同様であるということによって特徴づけられる。このようなビデオ番組のサマリを作成する際には、オーディオと画像の間の同期が重要である。したがって、同期はオーディオ中心型または画像中心型のいずれかであることが好ましい。

【0011】一実施例によれば、オーディオ中心型サマリ作成技術は、ビデオ番組に関連するオーディオ内容の重要な側面を確認する。必要な程度の同期を達成するため、画像サマリは、オーディオサマリを構成するオーディオセグメントに対応するビデオフレームを選択することによってのみ、生成される。画像中心サマリ作成技術は、まず、ビデオ番組の重要な画像セグメントを識別することによって画像トラックのサマリを作成する。その後、これらの重要なあるいは代表的な画像セグメントに対応するオーディオセグメントを、全体のビデオサマリに含める。

【0012】サマリを作成するプロセスは、画像、オーディオ信号、音声トランスクリプト、および字幕（クローズドキャプション）テキストからの手がかりおよび特徴を利用することによって容易化される。画像特徴、音声トランスクリプト、および字幕テキストは、オーディオサマリ作成を改善するために、対応するオーディオ特徴と組み合わされ、一方、オーディオ特徴、音声トランスクリプト、および字幕テキストは、よりよい画像サマリ作成を容易にするために、関連する画像特徴と組み合わされる。

【0013】オーディオ中心型、画像中心型、あるいはオーディオビジュアル統合型のサマリ作成を実現するため、以下では2つの実施例について説明する。1つの技術によれば、与えられたアプリケーションに対していずれのサマリ作成技術が好ましいかに応じて、あらかじめサマリ作成の選択（プレファレンス）を例示することが可能なトレーニングデータを用いて、機械学習が、ビデオ番組のオーディオあるいは画像トラックに適用される。この技術では、システムは、既知のアルゴリズム方式のうちの任意のものを用いて、サンプルビデオサマリ

に示される挙動を模倣し、このサンプルから、および、サンプルの固有のインプリメンテーションから、学習を行うことが可能である。必要な命令をシステムに提供するために、トレーニングデータが直ちに入手可能でない場合や容易に適用可能でない場合には、以下で説明するもう１つの実施例が、本発明の代替方法として、適用可能である。

【００１４】本発明の上記および関連するその他の利点は、添付図面を参照して、以下の好ましい実施例の詳細な説明を検討すれば、さらに明らかとなる。

【００１５】

【発明の実施の形態】図面を参照すると、図１は、機械学習によるビデオサマリ作成システム・方法に関する、本発明の一実施例の動作を示す流れ図である。図１を参照して、以下では、使用される数学的モデルのタイプ、オーディオおよびビジュアルサマリ作成に用いられる特徴、ならびに、オーディオおよびビジュアルサマリを整列させる方法について説明する。

【００１６】［機械学習フレームワーク］通常のビデオ番組は、オーディオトラックおよび画像トラックの両方を含み、これらはいずれも長く連続することがある。このようなビデオ番組のサマリを作成するには、そのビデオを構成するオーディオトラックおよび画像トラックの両方を、有意味かつ管理可能な操作ユニットに分節化しなければならない。例えば、有意味なオーディオ操作ユニットとしては、１個の単語、１個の句、１個の文、あるいはその他のコヒーレントな音響プロファイルを有するオーディオセグメントの発声がある。同様に、可能な画像操作ユニットの例には、単一のカメラショット、一連の連続するカメラショット、ある判断基準によってグループ分けされた画像フレームのクラスタなどがある。

【００１７】このような状況において、あるベクトルすなわち特徴セットＸで、オーディオまたは画像操作ユニットを表すことが可能である。さらに、Ｘは、いくつかの特徴ｘを含む。特徴ｘは、オーディオまたは画像操作ユニットに関連する画像特徴、オーディオ特徴、テキスト特徴（例えば、音声トランスクリプトや字幕からの重要なキーワード）とすることが可能である。ｎ個の特徴ｘが特定のベクトルすなわち特徴セットＸに存在する場合、Ｘ＝［x_1，x_2，…，x_n］である。サマリ作成作業は、与えられた特徴セットＸに対して、確率Ｐ（ｙ｜Ｘ）を計算する二分分類問題に変換される。ここでｙは２進（バイナリ）変数であり、その値１およびｏは、Ｘがサマリに含まれるか否かのそれぞれの状態を表す。この確率Ｐ（ｙ｜Ｘ）は、ルール（規則）によるアプローチを用いて決定することも可能であり、あるいは、機械学習法を用いて評価することも可能である。後者の場合、トレーニングデータが機械学習システムに提供され、システムは、提供されたトレーニングデータに従って、確率Ｐ（ｙ｜Ｘ）を予測するモデルを学習することになる。

【００１８】確率Ｐ（ｙ｜Ｘ）を評価するために、ナイーブベイズ法、決定木法、ニューラルネットワーク法、最大エントロピー法（これらには限定されない）などのような、既知のさまざまな機械学習技術のうちの任意のものを使用可能である。このような技術は、この技術分野の当業者に周知であるため、ここで詳細に説明する必要はない。

【００１９】［システム構成］上記のように、図１は、機械学習によるビデオサマリ要約作成システム・方法の一実施例の動作を示す概略流れ図である。システムは、ビデオ入力の画像トラックおよびオーディオトラックを検査する。さらに、システムは、入力ビデオに関連する字幕があればそれも検査することが可能である。ビデオサマリ作成システム・方法は、これらの３つの入力コンポーネント、すなわち、字幕、オーディオトラック、および画像トラックの間の整列を実行することが可能である。各入力コンポーネントに対する特徴抽出および特殊な操作も実行可能である。抽出された特徴および各コンポーネント操作の出力はその後、オーディオビジュアル統合型サマリ、または、オーディオ中心型サマリもしくは画像中心型サマリのいずれかを作成するために、機械学習によるサマリ作成モジュールに入力される。以下の操作が一般に、入力コンポーネントのそれぞれに関して実行される。

【００２０】サウンド（音）の検出と分類：　音楽、拍手、叫び声、爆発、雷鳴、銃声などのような非音声サウンドからなるオーディオセグメントを検出する。それらを、それぞれがコヒーレントな音響プロファイルを有するサウンドユニットに分節化する。これらのユニットを、それらの内容に従って分類する。各サウンドユニットに対して、以下のデータ、すなわち、オーディオトラック内でのそのサウンドユニットの開始時刻コード、そのサウンドユニットの継続時間、およびサウンドユニットのカテゴリあるいはタイプを出力する。

【００２１】音声認識：　サウンド検出・分類モジュールによって検出された非音声オーディオセグメントを取り除く。残りのオーディオセグメントに対して音声認識を実行して音声トランスクリプトを生成する。それぞれの認識語、オーディオトラック内でのその開始時刻コード、およびその継続時間を出力する。

【００２２】字幕と音声トランスクリプトの整列：　字幕と、音声認識器からの音声トランスクリプトとの間の整列を実行する。字幕は、タイピングミスを含むことがあり、音声認識器からの音声トランスクリプトは認識エラーを含むことがある。字幕と音声トランスクリプトの間の整列は、音声トランスクリプトの精度を改善するために有効である。

【００２３】音声ユニットと特徴ベクトル生成：　整列した音声トランスクリプトに基づいて音声操作ユニット

を生成し、各音声ユニットに対して特徴ベクトルを生成する。例えば、有意味な音声ユニットとしては、１個の単語、１個の句、１個の文、あるいはその他の有意味な音声内容を有するセグメントがある。

【００２４】音声ユニット重要度ランク付け：　各音声ユニットの重要度ランクを計算する。この重要度ランク付けは、例えば、米国特許仮出願第６０／２５４，５３５号（出願日：２０００年１２月１２日、発明の名称："Text Summarization UsingIR Technique And Singular Value Decomposition"）、および、米国特許出願第０９／８１７，５９１号（出願日：２００１年３月２６日、発明の名称："Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis"）に記載されているような、当業者に知られた方法を利用することが可能である（本出願人による特願２００１－３５６８１３号を参照）。

【００２５】話者ＩＤ検出：　音声ユニット・特徴ベクトル生成モジュールから音声ユニットを受け取る。それぞれの音声ユニットに対して、話者の識別を決定する（すなわち、話者識別検出）。

【００２６】画像セグメント分節化：　画像トラックを、それぞれがコヒーレントな画像プロファイルおよび動きプロファイルを有する個々の画像セグメントに分節化する。得られた画像セグメントは、画像操作ユニットとして使用可能である。

【００２７】画像特徴ベクトル生成：　画像特徴を抽出し、各画像セグメントに対して特徴ベクトルを形成する。特徴ベクトルを形成するためには、画像セグメント内容の何らかの側面を捕捉する任意の画像特徴が使用可能である。

【００２８】顔ＩＤ検出：　それぞれの画像セグメントに人間の顔が含まれていれば、それを検出し識別する。

【００２９】上記の操作が実行された後、出力は、機械学習によるサマリ作成モジュールに供給され、そこで、オーディオおよびビジュアルサマリが、前述のような機械学習フレームワークを用いて作成される。オーディオサマリ作成プロセスには、それぞれの音声あるいはサウンドユニットＸに対して、そのユニットがオーディオサマリに含まれるのに十分な重要性を有する確率Ｐ（ｙ｜Ｘ）を計算することが含まれる。上記のように、それぞれの音声あるいはサウンドユニットに関連する以下の特徴が、機械学習フレームワークで使用可能である。すなわち、その特徴とは、音声ユニットの開始時刻コード、継続時間、および重要度ランク、サウンドユニットの開始時刻コード、継続時間、およびカテゴリ、ならびに、対応する画像の顔識別、および画像特徴ベクトルである。それぞれの音声あるいはサウンドユニットＸに対する確率Ｐ（ｙ｜Ｘ）が計算された後、オーディオサマリがユーザ指定の長さL_{sum}に達するまで、確率Ｐ（ｙ｜Ｘ）の降順に音声ユニットあるいはサウンドユニットを選択することによって、オーディオサマリが作成される。

【００３０】他方、ビジュアルサマリ作成は、上記の操作で作成された画像セグメントを操作ユニットとして使用する。ビジュアルサマリ作成プロセスは、同様に、機械学習フレームワークを用いて、例えば各画像セグメントＳに対して、その画像セグメントがビジュアルサマリに含まれるのに十分な重要性を有する確率Ｐ（ｙ｜Ｓ）を計算する。上記のように、例えば、各画像セグメントＳに関連する以下の特徴が考えられる。すなわち、その特徴とは、長さ（すなわち、連続する、順次的な、あるいはその他の関連するフレームの個数）、画像特徴ベクトル、その画像セグメントに描画された人物あるいは顔の識別、黒フレームや画像ロゴなどのような特殊なフレームの存在、人間および物体（オブジェクト）の動き、ズームやパンなどのようなカメラの動き、対応する音声ユニットおよびサウンドユニット、ならびに、対応する音声ユニットに関連する話者の識別である。各画像セグメントＳに対する確率Ｐ（ｙ｜Ｓ）が計算された後、ビジュアルサマリがユーザ指定の長さL_{sum}に達するまで、確率Ｐ（ｙ｜Ｓ）の降順に画像ユニットを選択することによって、ビジュアルサマリが作成される。

【００３１】ビジュアルサマリは、必ずしも、それぞれの選択された画像セグメントを最初から最後まで含むことは必要でない。もとのビデオ番組を構成する画像セグメントの平均時間長は長いが、ユーザ指定のサマリ長L_{sum}は短い場合、ビジュアルサマリはほんの２、３個の画像セグメントによって構成されることになるため、もとのビジュアル内容の大幅な喪失につながる可能性がある。ユーザがビジュアルサマリ作成結果に影響を及ぼすことを可能にするため、ユーザは、サマリ長L_{sum}のみならず、最小再生時間L_{min}をも指定するように要求されることも可能である。L_{sum}は、全体のビジュアル内容を理解するためにユーザがどのくらいの長さの時間を使いたいかを示す一方、L_{min}は、幅指向ビジュアルサマリと深さ指向ビジュアルサマリの間の選択権をユーザに提供する。例えば、小さいL_{min}の値は、多数の短い画像セグメントからなる幅指向のビジュアルサマリを生成するために用いられる。他方、大きいL_{min}の値は、少数の長い画像セグメントからなる深さ指向のビジュアルサマリを生成するために用いられる。

【００３２】ユーザがL_{sum}およびL_{min}を指定した後、ビジュアルサマリ内に含まれることが可能な画像セグメントの総数は、Ｃ＝ｍｉｎ（L_{sum}／L_{min}，｜Ω｜）に等しい。ただし、｜Ω｜は、もとのビデオ内の画像セグメントの総数を表す。さらに、それぞれの選択された画像セグメントに割り当てられることが可能な時間長は、Ｌ＝L_{sum}／Ｃに等しい。この状況において、ビジュアルサマリは、確率Ｐ（ｙ｜Ｓ）の降順にＣ個の画像セグメントを選択し、そのＣ個の画像セグメントのそれぞれ

の最初のＬ秒間をとり、それらを時間の昇順に連結することによって作成される。

【００３３】［オーディオサマリとビジュアルサマリの間の整列］オーディオおよびビジュアルサマリが作成された後、解決すべき最後の問題は、どのようにしてこれらの２つのサマリを同期するかである。オーディオトラックＡおよび画像トラックＩからなるビデオシーケンスをＶ＝（Ｉ，Ａ）とする。Ｖのオーディオサマリは、$A_{sm}=\{A(t_i,\tau_i)\in A\mid i=1,\ldots,N(A_{sm})\}$と表される。ただし、Ａ（$t_i$，$\tau_i$）は、時刻 t_i に開始し時間 τ_i だけ継続するオーディオセグメントを表し、$N(A_{sm})$ は、A_{sm} を構成するオーディオセグメントの個数を表す。A_{sm} 内のすべてのオーディオセグメントは、それらの開始時刻 t_i の昇順に配列される。同様に、Ｖのビジュアルサマリは、$I_{sm}=\{I(t_i,\tau_i)\in I\mid j=1,\ldots,N(I_{sm})\}$と表され、すべてのコンポーネントはそれらの開始時刻の昇順にソートされる。

【００３４】上記のように、オーディオ中心型および画像中心型サマリは、同期の問題を最小にする。すなわち、同期は、単に、もとのビデオ番組から、画像またはオーディオのそれぞれの対応部分をとることによって実現可能である。オーディオ中心型サマリについては、Ａ（t_i，τ_i）$\in A_{sm}$ の場合、Ｉ（t_i，τ_i）$\in I_{sm}$ である。画像中心型サマリについては、Ｉ（t_i，τ_i）$\in I_{sm}$ の場合、Ａ（t_i，τ_i）$\in A_{sm}$ である。オーディオビジュアル統合型サマリを作成するためには、オーディオサマリとビジュアルサマリが機械学習フレームワークを用いて別個に作成されるため、それぞれのオーディオセグメントＡ（t_i，τ_i）$\in A_{sm}$ に対して、対応する画像セグメントＩ（t_i，τ_i）は必ずしも I_{sm} に属するとは限らず、逆も同様である。したがって、画像およびオーディオの両方の内容のカバレジを、それらのいずれをも犠牲にせずに最大化するため、オーディオサマリとビジュアルサマリの間でゆるい整列が実行される。

【００３５】オーディオビジュアル統合型サマリについては、どのオーディオ内容がどの画像内容と同期しなければならないか、およびその逆はどうかについての、システム設計者の、またはユーザの要求すなわちプレファレンスが、あらかじめ規定された整列指定として、サマリ作成システムに提供される。例えば、同期は、以下の場合に所望され、あるいは要求される。（１）ビジュアルサマリ内の画像セグメントが人物を示しており、対応するオーディオセグメント画素の人物の音声を含む場合、画像セグメントをそのオーディオ対応部分に、またはその逆に、同期することが所望される。（２）オーディオサマリ内のオーディオセグメントが爆発からなり、対応する画像セグメントが爆発を示している場合、オーディオセグメントをその画像対応部分に、またはその逆に、同期することが所望される。（３）オーディオセグメントが、ある有名人の名前に言及する音声を含み、その有名人の写真が、そのオーディオセグメントの小さい時間ウィンドウ内の画像セグメントに示されている場合、オーディオセグメントを、その有名人の写真を示す画像セグメントに、またはその逆に、同期することが所望される。

【００３６】一実施例によれば、オーディオビジュアル統合型サマリ作成は以下のように実行される。

【００３７】上記のビジュアルサマリ作成プロセスと同様に、オーディオビジュアル統合型サマリ作成は、２つのパラメータ、すなわち、ビジュアルサマリを構成する各画像セグメントに対するサマリ長 L_{sm} 、および最小再生時間 L_{min} を指定することをユーザに要求することによって開始される。ユーザが深さ指向ビジュアルサマリと幅指向ビジュアルサマリとの間の選択をすることを可能にすることとは別に、パラメータ L_{sm} を導入するもうＩつの目的は、オーディオサマリとビジュアルサマリの間の部分的整列を実現することである。整列の主な目標は、オーディオビジュアル統合型サマリがなめらかで自然に見えるようにし、もとのビデオのオーディオよびビジュアルの両方の内容のカバレジを、それらのいずれをも犠牲にすることなく、最大化することである。

【００３８】例えば、ニュース番組では、アナウンサーやレポーターによって話される文章は、ニュース記事の重要な内容を伝えている可能性が高く、オーディオサマリに含まれる高い確率が与えられる。このような文章の対応する画像部分は、スタジオのアナウンサーや現場のレポーターの接写である。オーディオサマリ内のそれぞれの話された文が、対応する画像部分とよく整列している場合、結果は、ほとんどアナウンサーやレポーターからなる画像部分を有するビデオサマリとなる。このようにして作成されるサマリは、自然でなめらかなものに見えるかもしれないが、このような自然さおよびなめらかさは、画像内容の相当な犠牲によりもたらされたものである。完全な整列により引き起こされるこの問題を解決するため、オーディオサマリとビジュアルサマリの間で、以下の部分的整列操作が代わりに実行される。

【００３９】Ｉ．上記のように、オーディオサマリは、確率の降順で、音声またはサウンドユニットを選択することによって作成される。

【００４０】２．オーディオサマリ内の各コンポーネントＡ（t_i，τ_i）に対して、対応する画像セグメントＩ（t_i，τ_i）の内容をチェックする。Ａ（t_i，τ_i）、Ｉ（t_i，τ_i）のペアが、システムに提供されたあらかじめ規定された整列要件を満たす場合、時刻 t_i から L_{sm} 秒間、Ａ（t_i，τ_i）をＩ（t_i，τ_i）と整列させる。そうでない場合は、Ａ（t_i，τ_i）に対して整列操作を実行しない。以下の記述において、時刻 t_i を整列点という。

【００４１】３．ステップ２で整列点が識別された後、

ビデオサマリ全体は、いくつかの時間パーティションに分割される。２つの隣り合う整列点t_i，t_{i+1}に対して、期間（t_i，t_{i+1}）に対するビジュアルサマリを作成するために、以下の操作を実行する。

【００４２】ａ．期間（t_i，t_i+L_{min}）を１（t_i，L_{min}）∈Ｉ（t_i，τ_i）で満たす。これは、Ａ（t_i，τ_i）と１（t_i，τ_i）の間の部分的整列を行う。

【００４３】ｂ．期間（t_i+L_{min}，t_{i+1}）に対するビジュアルサマリを作成するため、この期間に入る画像セグメントの集合Θを求める。この期間に含まれることが可能な画像セグメントの総数は、Ｃ＝ｍｉｎ（（$t_{i+1}-t_i-L_{min}$）／L_{min}，｜Θ｜）に等しい。ただし、｜Θ｜は、集合Θ内の画像セグメントの個数を表す。さらに、それぞれの画像セグメントに割り当てられることが可能な期間長は、Ｌ＝（$t_{i+1}-t_i-L_{min}$）／Ｃに等しい。Θから、最も高い確率を有するＣ個の画像セグメントを選択し、そのＣ個の画像セグメントのそれぞれの最初のＬ秒間をとり、それらを時間の昇順に連結することによって、この期間に対するビジュアルサマリを作成する。

【００４４】［ビデオサマリ作成の代替システム・方法］上記のように、機械学習フレームワークに基づくビデオサマリ作成のシステムおよび方法は、人間の専門家が前もって作成した十分な数のサンプルビデオサマリからなるトレーニングデータを必要とする。機械学習によるサマリ作成のシステムおよび方法は、専門家のサンプルビデオサマリから学習すること、および、サンプルビデオサマリに示される挙動を模倣することによってビデオサマリを作成することが可能である。しかし、場合によっては、専門家により作られたサンプルビデオサマリを得ることが高価すぎることや非常に困難なことがある。このような場合、トレーニングデータを必要としないシステムおよび方法を提供することが好ましい。

【００４５】図２は、トレーニングサンプルを必要としない代替的なビデオサマリ作成システム・方法の一実施例の動作を示す概略流れ図である。図２からわかるように、この代替システムは、前述の機械学習によるシステムのものと非常に類似した流れ図を有する。したがって、これから説明する代替システム・方法でも、第１実施例の場合と同様に、オーディオ中心型、画像中心型、またはオーディオビジュアル統合型のサマリを得ることが可能である。図２の流れ図において、この代替システムの、以下のモジュール以外はすべて、図１に示した対応するモジュールと同一である。

【００４６】ビジュアル内容による画像セグメントクラスタ化：　第１実施例と同様の画像セグメント分節化に加えて、画像セグメントを、それらのビジュアル類似度および動的レベルに基づいてクラスタ化する。このクラスタ化は、例えば、Ｙ．Gongand X. Liu, "Video Summarization Using Singular Value Decomposition", inProceedings of IEEE International Conference of Computer Vision and Pattern Recognition (CVPR'00)、に記載されているものや、Ｙ．Gong and X. Liu,"Summarizing Video By Minimizing Visual Content Redundancies", in Proceedings of IEEE International Conference of Multimedia and Expo (ICME'01)、に記載されているもののような方法を使用可能である。各フレームクラスタは、同じクラスタ内のすべての画像セグメントが互いに視覚的に類似しているような１個以上の画像セグメントからなる。

【００４７】すべての特徴抽出操作がそれぞれのモジュールによって実行された後、出力は、オーディオビジュアルサマリ作成モジュールに供給され、そこで、オーディオサマリもしくはビジュアルサマリのいずれか、またはオーディオビジュアル統合型サマリが以下で説明するように作成される。

【００４８】前述のシステムにおける機械学習によるビデオサマリ作成モジュールとは異なり、この場合のオーディオビジュアルサマリ作成モジュールは、それぞれの音声またはサウンドユニットＸに対する確率Ｐ（ｙ｜Ｘ）も、それぞれのフレームクラスタＳに対する確率Ｐ（ｙ｜Ｓ）も計算しない。代わりに、オーディオサマリがユーザ指定の長さL_{sum}に達するまで、音声ユニットを（音声ユニット重要度ランク付けモジュールから受け取った）その重要度ランクの降順に選択することによって、オーディオサマリを作成する。サウンドユニットは、例えば発見的ルールを用いて、ランク付けされ、オーディオサマリに含めるかどうか選択される。前述のように、音声ユニットの重要度ランク付けは、例えば、米国特許仮出願第６０／２５４，５３５号（出願日：２０００年１２月１２日、発明の名称："Text Summarization Using IR Technique And Singular Value Decomposition"）、および、米国特許出願第０９／８１７，５９１号（出願日：２００１年３月２６日、発明の名称："Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis"）に記載されているような、当業者に知られた方法を利用することが可能である（本出願人による特願２００１－３５６８１３号を参照）。さらに、重要度ランク付けは、発見的ルールと上記の方法との組合せを用いて決定することも可能である。例えば、このような発見的ルールは、重要な人物によって話された特定の語句や、爆発、自然災害、暴行などのような特別の事件を含む重要な画像セグメントに対応する音声ユニットに、より高いランクを与える。

【００４９】ビジュアルサマリを作成するため、代替システムもまた、２個のパラメータL_{sum}、L_{min}のユーザによる指定を必要とする。ここでも、各フレームクラスタＳの重要度をランク付けするために、発見的ルールが使用可能である。一般に、ビジュアル内容サマリ作成に対するシステム設計者の、またはユーザの知識、要望、

あるいはプレファレンスを反映する任意のルールが、発見的ルールとして使用可能である。このような発見的ルールは、特定の画像特徴や、有名人や、会社ロゴなどのような特別のフレームを含むフレームクラスタ、人間や物体の動きや、ズーム、パンなどのようなカメラの動きを有するフレームクラスタ、あるいは、対応する音声ユニットが重要であるか、重要な人物によって話されているフレームクラスタに、より高いランクを与える。

【００５０】さらに、各フレームクラスタの全時間長（構成する各画像セグメントの継続時間の和）もまた、ランク決定の過程で使用可能である。各フレームクラスタは複数の画像セグメントからなることがあるため、Ｉつのフレームクラスタが選択された後、ビジュアルサマリを作成するためには、そのクラスタ内のどの画像セグメントを使用すべきかを決定することが依然として必要である。画像セグメント選択の助けとなる情報が他にない場合、最も直接的な選択方法は、クラスタ内で最長の画像セグメントを選択することとなるであろう。この理由は、同じクラスタ内の画像セグメントはすべて視覚的に類似しているため、最長の画像セグメントは、最も完全なものであり、最もよくクラスタ全体を代表するからである。この状況において、ビジュアルサマリ内に含まれることが可能な画像セグメントの総数Ｃと、それぞれの選択された画像セグメントに割り当てられることが可能な時間長Ｌは、２個のパラメータL_{sum}、L_{min}を利用した前述のと同じ式を用いて計算可能である。ビジュアルサマリは、Ｃ個のフレームクラスタをそれらの重要度ランクの降順に選択し、そのＣ個のフレームクラスタのそれぞれから最長の画像セグメントの最初のＬ秒間をとった後、それらを時間の昇順に連結することによって、作成することができる。

【００５１】オーディオビジュアル統合型サマリについては、前述の機械学習によるシステムと同様に、代替システムもまた、どのオーディオ内容がどの画像内容と同期しなければならないかを示す整列指定と、パラメータL_{sum}、L_{min}のユーザによる入力とを必要とする。オーディオサマリが作成された後、オーディオサマリ内の各コンポーネントＡ（t_i，τ_i）は、その画像対応部分Ｉ（t_i，τ_i）とともに検査され、Ａ（t_i，τ_i）、Ｉ（t_i，τ_i）のペアがあらかじめ規定された整列要件を満たすかどうか調べられる。この検査は、ビデオサマリ全体をいくつかの時間パーティションに分割する整列点の集合を識別する。どのフレームクラスタが、および、選択されたフレームクラスタ内のどの画像セグメントが、どの時間パーティションを満たすために使用されるべきかを決定しなければならない。この整列操作は、以下の２つの主なステップからなる。

【００５２】１．オーディオサマリ内の各コンポーネントＡ（t_i，τ_i）に対して、対応する画像セグメントＩ（t_i，τ_i）の内容をチェックする。Ａ（t_i，τ_i）、Ｉ（t_i，τ_i）のペアが、あらかじめ規定された整列要件を満たす場合、時刻t_iからL_{sum}秒間、Ａ（t_i，τ_i）をＩ（t_i，τ_i）と整列させる。そうでない場合は、Ａ（t_i，τ_i）に対して整列操作を実行しない。ここで、時刻t_iを整列点という。

【００５３】２．ステップＩですべての整列点が識別された後、ビデオサマリ全体は、いくつかの時間パーティションに分割される。（ビジュアル内容によるフレームクラスタ化モジュールから得られる）クラスタ集合からのフレームクラスタを割り当てることによって、それぞれのパーティションを満たす。この割当ては、以下の２つの制約に適応しなければならない。

【００５４】ａ．単一割当て制約：　各フレームクラスタは、ただＩつの時間スロット割当てを受け取ることができる。

【００５５】ｂ．時間順序制約：　ビジュアルサマリを構成するすべての画像セグメントの時間順序は維持されなければならない。

【００５６】以下で、上記の整列操作のステップ２の実現法について説明する。ビデオサマリの全時間長L_{sum}が整列点によってＰ個のパーティションに分割され、パーティションｉの時間長がT_i（図３Ａ参照）であると仮定した場合、各時間スロットは少なくともL_{min}秒間の長さでなければならないため、パーティションｉは、

【数１】

$$S_i = \lfloor T_i / L_{\min} \rfloor$$

個の時間スロットを提供することが可能であり、したがって利用可能な時間スロットの総数は$S_{total} = \Sigma_{i=1}^{P} S_i$となる。ここで、問題は次のようになる。ビデオサマリの全部でＯ個のフレームクラスタとS_{total}個の時間スロットが与えられた場合に、上記の２つの制約を満たすように、フレームクラスタと時間スロットの間の最適なマッチングを決定せよ。

【００５７】若干の再定式化によって、今述べた問題を、最大２部マッチング問題に変換することができる。頂点の有限集合をＶとし、Ｖ上の辺集合をＥとする無向グラフをＧ＝（Ｖ，Ｅ）で表す。２部グラフとは、無向グラフＧ＝（Ｖ，Ｅ）であって、Ｖが、（ｕ，ｖ）∈Ｅならばｕ∈Ｌかつｖ∈Ｒかつｕ∈Ｒかつｖ∈Ｌかのいずれかが成り立つような２つの集合ＬおよびＲに分割可能であるようなもののことである。すなわち、すべての辺は、２つの集合ＬとＲの間をつなぐ。マッチングとは、辺の部分集合Ｍ⊆Ｅであって、ｕ∈Ｌかつｖ∈Ｒである任意の頂点対（ｕ，ｖ）に対して、Ｍの高々Ｉつの辺がｕとｖの間を連結するようなもののことである。

【００５８】最大マッチングとは、マッチングＭであって、任意のマッチングM'に対して、$|M| \geq |M'|$となるようなもののことである。この問題に最大２部マ

ッチングを適用するため、各頂点ｕ∈Ｌを用いてフレームクラスタを表し、各頂点ｖ∈Ｒを用いて時間スロットを表す。辺（ｕ，ｖ）が存在するのは、フレームクラスタｕが、時間順序制約に違反せずに時間スロットｖをとることができる場合である。フレームクラスタが、もとのビデオの前半からのものと、もとのビデオの後半からのものという複数の画像セグメントからなる場合、このフレームクラスタは、それから出てＲ内の相異なる頂点に至る複数の辺を有することになる。

【００５９】最大２部マッチング解は、すべてのフレームクラスタと時間スロットの間の最適割当てである。なお、最適割当ては必ずしも一意的であるとは限らない。

【００６０】図３Ａに、オーディオサマリとビジュアルサマリの間の整列プロセスを示す。この図において、もとのビデオ番組は７０秒間の長さであり、その画像トラックは、それぞれ１０秒間継続する７個の画像セグメントからなり、オーディオトラックは、それぞれ長さ１０秒間の７個の話された文からなる。ユーザは、L_{sum} ＝２０秒、およびL_{win} ＝３秒と設定している。オーディオサマリ作成は、２個の話された文Ａ（０，１０）およびＡ（３０，１０）を選択し、ビジュアル内容によるクラスタ化は、次の５個のクラスタを生成したと仮定する：

Ｉ（０，１０）からなるクラスタ１、

Ｉ（１０，１０）およびＩ（５０，１０）からなるクラスタ２、

Ｉ（３０，１０）からなるクラスタ３、

Ｉ（２０，１０）およびＩ（４０，１０）からなるクラスタ４、

Ｉ（６０，１０）からなるクラスタ５。

【００６１】オーディオサマリがＡ（０，１０）およびＡ（３０，１０）から形成されているので、対応する画像セグメントＩ（０，１０）およびＩ（３０，１０）の内容を検査し、Ａ（０，１０）およびＡ（３０，１０）に対して整列操作が要求されるかどうかを判定する必要がある。Ｉ（０，１０）およびＩ（３０，１０）は話された文Ａ（０，１０）、Ａ（３０，１０）のそれぞれの話者を表示していると仮定する。その場合、整列ルールにより、L_{win}（３）秒間、Ｉ（０，１０）はＡ（０，１０）と整列し、Ｉ（３０，１０）はＡ（３０，１０）と整列することになる。Ｉ（０，１０）およびＩ（３０，１０）は、一度使用されたため、これらはビジュアルサマリの他の部分で使用されることはない。

【００６２】これらの２つの整列点により、ビジュアルサマリの残りの期間は２つのパーティションに分割される。各パーティションは、高々２個の時間スロットを提供することが可能な７秒間継続する。整列のために３個のフレームクラスタおよび４個の時間スロットが残っているため、この整列作業に対して、図３Ｂに示す２部グラフがある。フレームクラスタ２は、２個の画像セグメント１（１０，１０）および１（５０，１０）からなるため、パーティション１またはパーティション２のいずれに時間スロットをとることも可能である。Ｉ（１０，１０）がフレームクラスタ２から選択される場合、これはパーティション１に時間スロット２または３のいずれかをとることができる。他方、Ｉ（５０，１０）が選択される場合、これはパーティション２に時間スロット５または６のいずれかをとることができる。したがって、クラスタ２から出る４本の辺、すなわち、時間スロット２への辺、時間スロット３への辺、時間スロット５への辺、および時間スロット６への辺が存在する。同様に、クラスタ４から出る４本の辺、すなわち、時間スロット２への辺、時間スロット３への辺、時間スロット５への辺、および時間スロット６への辺が存在する。

【００６３】他方、フレームクラスタ５は、ただ１つの画像セグメントＩ（６０，１０）からなり、パーティション２に時間スロット５または６のいずれかをとることができる。したがって、フレームクラスタ５から出る２本の辺が存在する。

【００６４】図３Ｂの２部グラフに対してはいくつかの可能な最大マッチング解が存在する。図４Ａおよび図４Ｂは２つのそれぞれの解を示す。図４Ａに示す解（ｉ）では、時間スロット３が未割当てのままである。図４Ｂに示す解（ｉｉ）では、時間スロット５が未割当てのままである。この場合、すべてのフレームクラスタが使用されているため、複数の画像セグメントを有するフレームクラスタを用いて、空き時間スロットを満たす必要がある。解（ｉ）（図４Ａ）の場合、フレームクラスタ４の画像セグメントＩ（２０，１０）が、空き時間スロットを満たすために使用されなければならない。解（ｉｉ）（図４Ｂ）の場合、フレームクラスタ２の画像セグメント１（５０，１０）が、空き時間スロットを満たすために使用されなければならない。

【００６５】上記の例は次のことを例示している。すなわち、最大２部マッチングは、利用可能なフレームクラスタと時間スロットの間の最適なマッチングを求めるが、特に、利用可能なフレームクラスタの数より多くの利用可能な時間スロットがあるときには、一部の時間スロットを未割当てのまま残すことがある。これらの未割当て時間スロットを満たすために、単一割当て制約をゆるめ、複数の画像セグメントを有するフレームクラスタを検査し、まだ使用されていない適当なセグメントを選択することが可能である。このようにして、時間順序制約は満たされる。ゆるめられた単一割当て制約に対するそれぞれの解を図５Ａおよび図５Ｂに示す。

【００６６】なお、最大２部マッチング操作は、不正な解を生成することがある。図６Ａおよび図６Ｂは、これの２つの例を示す。例（ｉ）（図６Ａ）では、画像セグメントＩ（６０，１０）が画像セグメントＩ（５０，１０）の前に置かれているため、時間順序制約に違反して

いる。例（ｉｉ）（図６Ｂ）では、割当てはいずれの制約にも違反していないが、１（２０，１０）を時間スロット２に割り当てることが、時間スロット３の割当てを不可能にしている。しかし、これらの不正な解は、これらを２つの制約に照らして検査することによって容易に検出され、各パーティションにおいて時間スロットに割り当てられる画像セグメントを並べ替えることによって補正することができる。例（ｉ）（図６Ａ）の場合、問題は、パーティション２に割り当てられた２個の画像セグメントを時間の昇順にソートすることによって補正することができる。例（ｉｉ）（図６Ｂ）の場合、まず、フレームクラスタ２からの画像セグメントＩ（１０，１０）（これは、パーティション１に割り当てられることが可能な唯一の残りの画像セグメントである）を用いて空き時間スロットを満たした後に、そのパーティション内の２個の画像セグメントをソートすることによって、最終的な解に達することができる。

【００６７】まとめると、整列操作のステップ２は、次のように記述することができる。

【００６８】１．整列点が識別された後、割当てのために残っているフレームクラスタおよび時間スロットの個数を決定し、それに応じて２部グラフを作る。

【００６９】２．最大２部マッチングアルゴリズムを適用して可能な解を求める。

【００７０】３．解を２つの制約について検査し、各パーティション内の画像セグメントをソートすることによって違反を補正する。

【００７１】４．未割当て時間スロットが存在する場合、単一割当て制約をゆるめ、複数の画像セグメントを有するフレームクラスタを検査し、まだ使用されていないセグメントで時間順序制約を満たす適当なセグメント＊を選択する。

【００７２】以上、好ましい実施例を参照して、本発明について詳細に説明したが、本発明の技術的範囲および技術思想の範囲内のさまざまな変形は、この技術分野の当業者には明らかである。したがって、本発明は、特許請求の範囲の技術的範囲によってのみ限定されるとみなされるべきである。

【００７３】

【発明の効果】以上詳細に説明したように、本発明によれば、オーディオと画像の内容の厳密な同期が要求されないときには、オーディオビジュアル統合型サマリ作成技術を用い、オーディオ内容と画像内容の同期を要求するビデオ番組の場合には、オーディオ中心型または画像中心型のいずれかの方法を用いてサマリが作成される。これにより、入力ビデオから抽出された画像、オーディオ、およびテキスト特徴をシームレスに統合し、オーディオ中心型、画像中心型、およびオーディオビジュアル統合型の高品質のサマリを作成することができる。

【図面の簡単な説明】

【図１】本発明のオーディオビジュアルサマリ作成システム・方法の一実施例の動作を示す流れ図である。

【図２】本発明のオーディオビジュアルサマリ作成システム・方法の代替実施例の動作を説明する流れ図である。

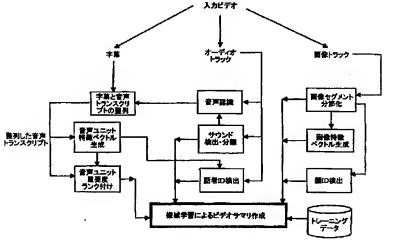
【図３】Ａは、オーディオサマリとビジュアルサマリの間の整列プロセスを示す図である。Ｂは、その整列のためのフレームワークを示す図である。

【図４】時間順序制約を満たす代替解を示す図である。

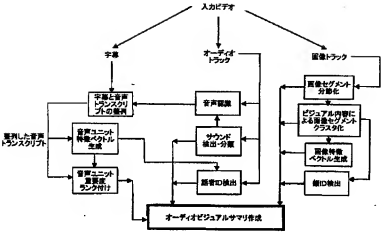
【図５】時間順序制約を満たす代替解を示す図である。

【図６】本発明の方法から得られる不正な解を示す図である。

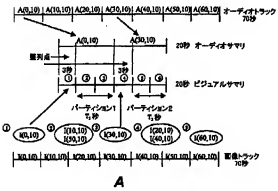
【図１】

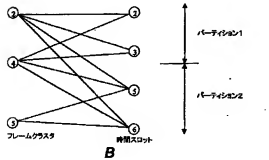


【図２】

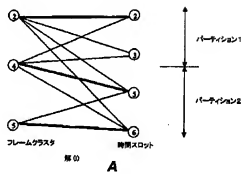

入力ビデオ
字幕
オーディオトラック
画像トラック
字幕セグメント分割化
字幕と音声トランスクリプトの整列
音声認識
ビジュアル内容による音響セグメントクラス分化
整列した音声トランスクリプト
音声ユニット特徴ベクトル生成
サウンド検出・分類
音声ユニット重要度ランク付け
画像特徴ベクトル生成
話者ID検出
顔ID検出
オーディオビジュアルサマリ作成

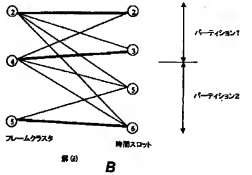
【図３】


A(0,10)
A(10,10)
A(20,10)
A(30,10)
A(40,10)
A(50,10)
A(60,10)
オーディオトラック 70秒
20秒 オーディオサマリ
重複区間
3秒
20秒 ビジュアルサマリ
パーティション1 T1秒
パーティション2 T2秒
I(0,10)
I(10,10)
I(20,10)
I(30,10)
I(40,10)
I(50,10)
I(60,10)
画像トラック 70秒
A

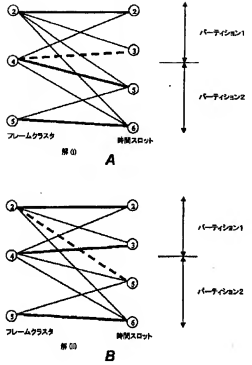

パーティション1
パーティション2
フレームクラスタ
時間スロット
B

【図４】

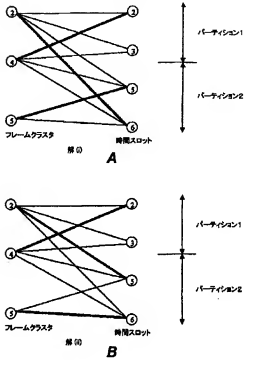

パーティション1
パーティション2
フレームクラスタ
時間スロット
顔 (i)
A


パーティション1
パーティション2
フレームクラスタ
時間スロット
顔 (j)
B

【図５】



【図６】



フロントページの続き

(51) Int.Cl.'	識別記号	FI		ﾃｰﾏｺｰﾄﾞ(参考)
H04N 5/91		H04N 5/91	N C	

(72)発明者　シン　リュウ
アメリカ合衆国，ニュージャージー　08540　プリンストン，４　インディペンデンス　ウエイ，エヌ・イー・シー・ユー・エス・エー　インク内

Fターム(参考)　5C053 FA14 GA16 GB05 JA01
5D015 AA03 AA06 FF00 GG00 HH00
KK02 LL11